

Rapid Detection of the ACMG/ACOG-Recommended 23 *CFTR* Disease-Causing Mutations Using Ion Torrent Semiconductor Sequencing

Aaron M. Elliott,* Joy Radecki, Bellal Moghis, Xiang Li, and Anja Kammesheidt

Ambry Genetics, Aliso Viejo, California 92656, USA

Cystic fibrosis (CF) is one of the most frequently diagnosed autosomal-recessive diseases in the Caucasian population. For general-population CF carrier screening, the American College of Medical Genetics (ACMG)/American College of Obstetricians and Gynecologists (ACOG) have recommended a core panel of 23 mutations that will identify 49–98% of carriers, depending on ethnic background. Using a genotyping technology that can rapidly identify disease-causing mutations is important for high-throughput general-population carrier screening, confirming clinical diagnosis, determining treatment options, and prenatal diagnosis. Here, we describe a proof-of-concept study to determine whether the Ion Torrent Personal Genome Machine (PGM) sequencer platform can reliably identify all ACMG/ACOG 23 CF transmembrane conductance regulator (*CFTR*) mutations. A WT CF specimen along with mutant DNA specimens representing all 23 *CFTR* mutations were sequenced bidirectionally on the Ion Torrent 314 chip to determine the accuracy of the PGM for *CFTR* variant detection. We were able to reliably identify all of the targeted mutations except for 2184delA, which lies in a difficult, 7-mer homopolymer tract. Based on our study, we believe PGM sequencing may be a suitable technology for identifying *CFTR* mutations in the future. However, as a result of the elevated rate of base-calling errors within homopolymer stretches, mutations within such regions currently need to be evaluated carefully using an alternative method.

KEY WORDS: DNA, high-throughput DNA sequencing, next-generation sequencing, sequence analysis

INTRODUCTION

The emergence of next-generation sequencing (NGS) technologies has provided new and reliable approaches to diagnostic testing. These rapidly evolving technologies have demonstrated advantages over Sanger sequencing by capillary electrophoresis, such as the abilities to generate megabases (Mb) to gigabases of data and to detect genetic mosaicism.^{1,2} However, the current NGS platforms have several weaknesses, including sequencing time, sample scalability, and cost of entry, which need to be addressed if these technologies are going to be used for routine diagnostic purposes. Moreover, the total amount of data produced is typically excessive when sequencing a small number of genes.³ A new sequencing technology, developed by Ion Torrent (Guilford, CT, USA; now owned by Life Technologies, Carlsbad, CA, USA), has solved many of these challenges.^{4,5} With the use of semiconductor sequencing, the Ion Torrent Personal Genome Machine (PGM) can currently generate >~10 Mb pairs (Mbp) of sequence data on the first-generation 314 chip, within several hours of ma-

chine run time. To test the feasibility of using the PGM process for clinical genotyping, we assessed the performance to detect common mutations in the cystic fibrosis transmembrane conductance regulator (*CFTR*) gene responsible for CF.

CF is one of the most common genetic diseases affecting Caucasian individuals. More than 1800 mutations have been documented in the *CFTR* gene, however only a small subset accounts for the majority of CF cases.⁶ For general-population screening, the American College of Medical Genetics (ACMG) and the American College of Obstetricians and Gynecologists (ACOG) have recommended a panel of 23 mutations, which will detect ~88% of carriers in non-Hispanic Caucasians.⁷ In the current proof-of-concept study, we sequenced previously characterized CF patients' DNA to determine whether the Ion Torrent PGM can reliably detect the recommended ACMG/ACOG *CFTR* mutations.

MATERIALS AND METHODS

Genomic DNA Preparation and PCR Amplification

Mutation analysis was conducted on previously characterized, archived genomic DNA. Upon submitting blood

*ADDRESS CORRESPONDENCE TO: Aaron M. Elliott, Ambry Genetics, 15 Argonaut, Aliso Viejo, CA 92656, USA. E-mail: aelliott@ambrygen.com
doi: 10.7171/jbt.12-2301-003

TABLE 1

Data Generation from Three PGM Runs				
Run	Total number of reads	Total bases (Mbp)	AQ17 total bases (Mbp)	AQ17 avg. read length
CF WT	101,211	8.5	6.5	68
CF 23 pooled mutants	222,247	18.6	12.52	64
CF mutant	135,000	11.7	8.8	72

samples for CFTR diagnostic testing, all individuals in the study provided written, informed consent for further research testing. DNA had been isolated from peripheral blood using the DNeasy Blood & Tissue Kit (Qiagen, Valencia, CA, USA). To specifically target the 23 ACMG/ACOG variants, 21 separate PCR primer pairs were designed (Integrated DNA Technologies, San Diego, CA, USA) and used to amplify mutant loci in 23 individual patients. Each patient sample was amplified only with the primer pair covering the mutation of interest. WT sample was processed with the same 21 primer pairs. In accordance with Ion Torrent recommendations, amplicon sizes ranged from 76 to 133 bp. Amplifications were performed in a Bio-Rad MyCycler (Bio-Rad, Hercules, CA, USA) and

consisted of 1 × Qiagen HotStarTaq Master Mix (Qiagen), 0.5 μM forward primer, 0.5 μM reverse primer, and 50 ng genomic DNA. PCR amplifications were performed with the following conditions: 95°C for 15 min, followed by a touchdown program of 94°C for 15 s, 60°C for 30 s, and 72°C for 15 s, with a 0.5°C decrease of the annealing temperature every cycle for eight cycles. After completion of the touchdown program, 30 additional cycles were subsequently performed (94°C for 15 s, 55°C for 30 s, and 72°C for 15 s), ending with a 10-min extension at 72°C. Amplicon concentration was determined using an Agilent BioAnalyzer DNA 1000 LabChip (Agilent Technologies, Santa Clara, CA, USA). Amplicons were then pooled together in equimolar concentrations and purified using the

TABLE 2

CFTR Variant Coverage, Mutant Read Percentage, and Base-Call Accuracy from a WT Library Using PGM Sequencing				
Variant	cDNA position	Coverage	Mutant read %	Accuracy/base
G85E	c.254G > A	408	0	99.5
R117H	c.350G > A	3627	0	99.9
621 + 1G > T	c.489 + 1G > T	245	0	99.6
711 + 1G > T	c.579 + 1G > T	2660	0	99.9
R334W	c.1000C > T	5419	0	99.7
R347P	c.1040G > C	3562	0	99.4
A455E	c.1364C > A	10,340	0	99.9
ΔI507	c.1519_1521delATC	6507	0	98.6
ΔF508	c.1521_1523delCTT	6507	0	99.4
1717-1G > A	c.1585-1G > A	2086	0	99.2
G542X	c.1624G > T	854	0	97.8
G551D	c.1652G > A	3901	0	99
R553X	c.1657C > T	3915	0	99.9
R560T	c.1679G > C	3924	0	99.6
1898 + 1G > A	c.1766 + 1G > A	1793	0	97.6
2184delA ^a	c.2052delA	2001	35%	63.6
2789 + 5G > A	c.2657 + 5G > A	293	0	100
3120 + 1G > A	c.2988 + 1G > A	2408	0	100
R1162X	c.3484C > T	9610	0	98.1
3659delC	c.3528delC	9271	0	100
3849 + 10kbC > T	c.3717 + 12191C > T	10,157	0	99.9
W1282X	c.3846G > A	4789	0	95.6
N1303K	c.3909C > G	3236	0	99.5

^aThe 2184delA variant lies in a homopolymer stretch of seven adenines and is not detected accurately as a result of homopolymer-length sequencing errors.

Qiagen MinElute PCR Purification Kit (Qiagen), according to the manufacturer's instructions.

Ion Torrent PGM Library Preparation and Sequencing

An Ion Torrent adapter-ligated library was made following the manufacturer's Ion Fragment Library Kit (Life Technologies) protocol (Part #4467320 Rev. A). Briefly, 50 ng pooled amplicons were end-repaired, and Ion Torrent adapters P1 and A were ligated using DNA ligase. Following AMPure bead (Beckman Coulter, Brea, CA, USA) purification, adapter-ligated products were nick-translated and PCR-amplified for a total of 10 cycles. The resulting library was purified using AMPure beads (Beckman Coulter) and the concentration and size determined using an Agilent BioAnalyzer DNA High-Sensitivity LabChip (Agilent Technologies). Sample emulsion PCR, emulsion breaking, and enrichment were performed using the Ion Xpress Template Kit (Part #4467389 Rev. B), according to the manufacturer's instructions. Briefly, an input concentration of one DNA template copy/Ion Sphere Particles (ISPs) was added to the emulsion PCR master mix and the emulsion generated using an IKA DT-20 mixer (Life Technologies). Next, ISPs were recovered and template-positive ISPs enriched for using Dynabeads MyOne Streptavidin C1 beads (Life Technologies). ISP enrichment was confirmed using the Qubit 2.0 fluorometer (Life Technologies), and the sample was prepared for sequencing using the Ion Sequencing Kit protocol (Part #4467391 Rev. B). The complete sample was loaded on an Ion 314 chip and sequenced on the PGM for 65 cycles.

Bioinformatic Analysis

Data from the PGM runs were processed initially using the Ion Torrent platform-specific pipeline software Torrent Suite v1.3.1 to generate sequence reads, trim adapter sequences, filter, and remove poor signal-profile reads. Generated sequence files were aligned to the *CFTR* genomic sequence (NC_000007.13), and the targeted variants were detected using a custom bioinformatics pipeline, developed by Ambry Genetics for a *CFTR* screening panel (CF102). In this pipeline, International Union of Biochemistry (IUB) ambiguity codes are used to represent all Reference single nucleotide polymorphisms (SNP) alleles at known (dbSNP132) loci. Alignment between an IUB ambiguity code and any one of the SNP bases must produce a perfect match. For variant detection, a minimum coverage of 20 must be achieved, and at least 5% of reads must represent the targeted mutation. All variant calls were also confirmed using SoftGenetics NextGENe V2.14 Ion Torrent software module (SoftGenetics, State College, PA, USA) with parameters of 80% minimum read match to reference sequence and 30 bp match size. A $\geq 5\%$ mutant read coverage was selected for variant identification.

RESULTS

To investigate whether the PGM is suitable for detecting the 23 *CFTR* mutations composing the ACMG/ACOG screening panel, we first prepared a PCR amplicon library from a WT individual characterized previously. For this run, the PGM 314 chip output was 8.5 Mbp, with $\sim 76\%$ aligning to the *CFTR* reference sequence at AQ17 (one

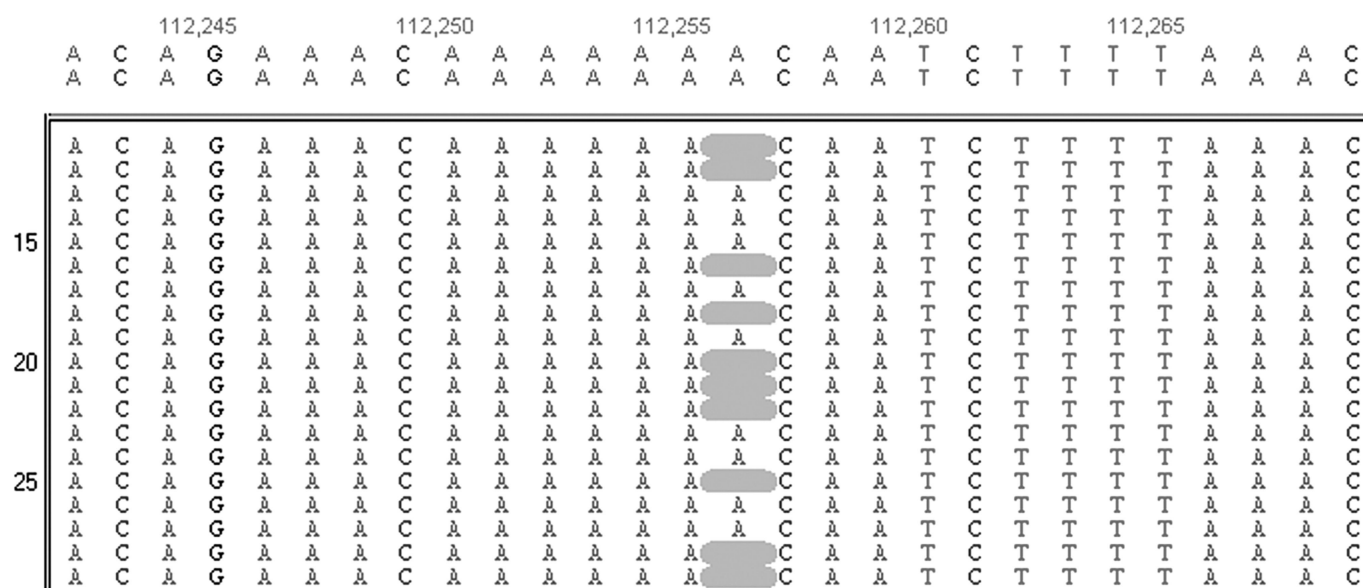


FIGURE 1

PGM sequence analysis displaying the false-deletion call at the 2184delA locus in a WT sample using SoftGenetics software.

error in every 50 bp; Table 1). A custom alignment script designed to target the *CFTR* mutation loci was used for variant identification. All variant calls were additionally confirmed using the SoftGenetics NextGENe Ion Torrent software module. Mean and median coverage for the 23 selected genomic positions was 4239 and 3627, respectively. The correct WT sequence for *CFTR* was observed for 22 of the 23 variant positions (Table 2). A false-positive 2184delA call was made in 35% of the reads spanning this variant, with no bias in directionality (Fig. 1). The 2184delA variant lies in a homopolymer stretch of seven adenines and was not detected accurately as a result of homopolymer-length sequencing errors. Using the same target amplicon primer set, we had previously been able to accurately identify the correct WT call at the 2184delA position in over 99% of reads using the Illumina Genome Analyzer IIx (GAIIx; data not shown). To determine the accuracy of the correct WT base call at each investigated position, we calculated the percentage ratio of the base consistent with the WT position to the total number of aligned reads. Accuracy call rates were >95% for 22 of the variants, and 17 of the 22 variants had call accuracy >99% (Table 2). The 2184delA

error rate was high, with an accuracy of only 63.6%. Similar results were observed in repeat runs.

Next, we evaluated whether PGM sequencing could correctly identify the actual *CFTR* mutations for all 23 loci. A PGM library was constructed, which contained 23 CF patient DNAs, each representing one of the 23 ACMG/ACOG mutations. All specimens were heterozygous variants, except for the $\Delta F508$ specimen, which was homozygous. As a result of the fact that some of the variants are in close proximity or share the same primer pair, the predicted “mutant-to-WT read ratios” for a heterozygous or homozygous variant were adjusted during analysis. For example, mutants $\Delta I507$ and $\Delta F508$ share the same primer pair covering the regions of interest. Therefore, the mutant-to-WT ratio for a $\Delta F508$ homozygous mutant will be 50% instead of 100%. Likewise, the $\Delta I507$, which is a heterozygous sample, will be 25% instead of 50%. All mutations had been identified initially using Sanger sequencing and had also been confirmed on the Illumina GAIIx platform using the identical amplicon designs as used in the current PGM experiments. For this data set, the PGM 314 chip output was 18.6 Mbp, with ~67% aligning to the *CFTR*

TABLE 3

PGM *CFTR* Variant Coverage and Mutant Read Percentage from a Pooled Mutant Library Representing All 23 ACMG/ACOG Mutations

Variant	cDNA position	Coverage	Mutant read %	Predicted read %	Genotype
G85E	c.254G > A	93	33	50	Het
R117H	c.350G > A	6228	39	50	Het
621 + 1G > T	c.489 + 1G > T	1243	46	50	Het
711 + 1G > T	c.579 + 1G > T	1352	29	50	Het
R334W	c.1000C > T	13,284	8	25	Het
R347P	c.1040G > C	9454	27	25	Het
A455E	c.1364C > A	19,527	43	50	Het
$\Delta I507$	c.1519_1521delATC	15,587	14	25	Het
$\Delta F508$	c.1521_1523delCTT	15,587	68	50	Homo
1717-1G > A	c.1585-1G > A	3584	36	50	Het
G542X	c.1624G > T	610	41	50	Het
G551D	c.1652G > A	6714	16	17	Het
R553X	c.1657C > T	6670	15	17	Het
R560T	c.1679G > C	6395	22	17	Het
1898 + 1G > A	c.1766 + 1G > A	3293	49	50	Het
2184delA ^a	c.2052delA	2256	63	50	Het
2789 + 5G > A	c.2657 + 5G > A	1765	54	50	Het
3120 + 1G > A	c.2988 + 1G > A	7447	40	50	Het
R1162X	c.3484C > T	19,060	54	50	Het
3659delC	c.3528delC	28,321	30	50	Het
3849 + 10kbC > T	c.3717 + 12191C > T	27,102	46	50	Het
W1282X	c.3846G > A	9219	48	50	Het
N1303K	c.3909C > G	4842	49	50	Het

^aThe 2184delA variant lies in a homopolymer stretch of seven adenines and is not accurately detected as a result of homopolymer-length sequencing errors.

reference sequence at AQ17 (Table 1). Analysis of the PGM data identified all 23 *CFTR* mutations with a mean and median read depth of 9114 and 6670, respectively (Table 3). Variant G85E had relatively low coverage of 93 reads. The G85E region consistently resulted in low read counts for all PGM runs that we conducted. The variant position is flanked by six thymines in the forward position and five thymines in the reverse, making it a difficult region to generate quality sequencing data. Low read counts in relation to the other variants were also observed for G85E using the Illumina GAIIX platform. Similar to what was observed in the WT data set, the 2184delA mutant read distribution was higher (63%) than the predicted value (50%) as a result of sequencing errors in the homopolymer stretch, resulting in erroneous single base-pair deletion calls. In addition, the R334W variant had 8% mutant reads compared with the predicted value of 25%. Coverage of the R334W variant was high, with 13,284 reads, and after inspecting the run data, we could not identify any particular reason why the mutant read distribution was skewed. A repeat run using the same library resulted in 19% R334W mutant reads, illustrating that there is some degree of run

variability. The remaining variants were in line with the predicted mutant-to-WT read ratio distributions.

Finally, we sought to analyze the PGM process to detect the correct *CFTR* mutations in a diagnostic setting. A PGM library, generated from an individual harboring two previously identified, disease-causing *CFTR* mutations, was constructed and sequenced on the PGM. The total data output for the run was 11.7 Mbp, with ~75% aligning to the *CFTR* reference sequence at AQ17 (Table 1). The mean coverage for the 23 *CFTR* variant positions was 5573 with a median coverage of 4545. Analysis of the data correctly identified the two heterozygous mutations $\Delta F508$ and G542X, with mutant read distributions of 47% and 41%, respectively (Table 4 and Fig. 2). Consistent with previous runs, there was a high percentage of false-positive 2184delA reads.

DISCUSSION

The purpose of this study was to analyze the performance of the current Ion Torrent PGM process to determine if it is readily suitable for diagnostic sequencing. Using samples characterized previously, we analyzed the PGM's data out-

TABLE 4

PGM *CFTR* Variant Coverage and Mutant Read Percentage from an Individual Harboring Two Disease-Causing *CFTR* Mutations

Variant	cDNA position	Coverage	Mutant read %
G85E	c.254G > A	237	0
R117H	c.350G > A	3774	0
621 + 1G > T	c.489 + 1G > T	936	0
711 + 1G > T	c.579 + 1G > T	2018	0
R334W	c.1000C > T	10,899	0
R347P	c.1040G > C	7720	0
A455E	c.1364C > A	14,525	0
$\Delta I507$	c.1519_1521delATC	8855	0
$\Delta F508$	c.1521_1523delCTT	8855	47
1717-1G > A	c.1585-1G > A	2216	0
G542X	c.1624G > T	2035	41
G551D	c.1652G > A	4581	0
R553X	c.1657C > T	4545	0
R560T	c.1679G > C	4774	0
1898 + 1G > A	c.1766 + 1G > A	2702	0
2184delA ^a	c.2052delA	2837	18.5
2789 + 5G > A	c.2657 + 5G > A	860	0
3120 + 1G > A	c.2988 + 1G > A	4347	0
R1162X	c.3484C > T	12,039	0
3659delC	c.3528delC	7169	0
3849 + 10kbC > T	c.3717 + 12191C > T	11,588	0
W1282X	c.3846G > A	6187	0
N1303K	c.3909C > G	4479	0

^aThe 2184delA variant lies in a homopolymer stretch of seven adenines and is not accurately detected as a result of homopolymer-length sequencing errors.

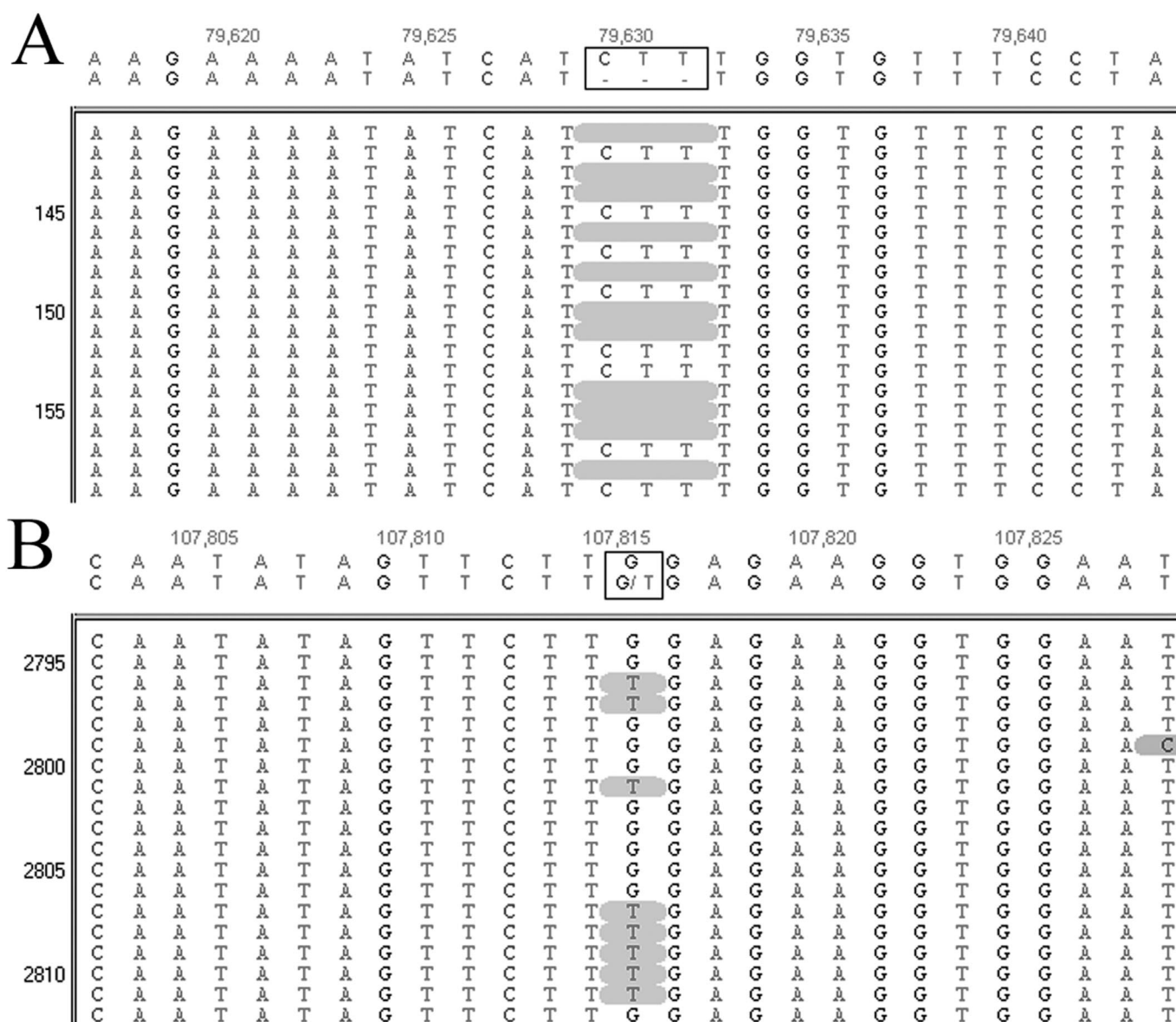


FIGURE 2

SoftGenetics PGM sequence analysis illustrating the *CFTR* mutations Δ F508 (A) and G542X (B) in a single CF sample. A Thymine/Cytosine nucleotide incorporation error can be detected in Line 6 of the sequence alignment (B).

put to correctly detect the 23 *CFTR* mutations comprising the ACMG/ACOG screening panel. Sequencing the WT *CFTR* library enabled us to immediately identify that homopolymer stretches were the main limitation of the PGM technology. Although WT at the 2184delA position, sequencing errors resulted in 35% mutant reads for this variant. The high error rate for this seven-adenine homopolymer tract was consistent in other sequencing runs. In most cases, the sequencing error type had one less base than the reference sequence, resulting in a false-deletion mutant call. The *CFTR* genomic sequence is highly complex and contains numerous stretches of homopolymers throughout the exonic regions. We routinely observed homopolymer

sequencing errors in tracts greater than five nucleotides throughout the amplified regions. The inability to distinguish true sequence variants from sequencing errors in homopolymer stretches is of critical importance if this technology is used for diagnostic-targeted resequencing. It is widely known that homopolymer sequencing errors are limitations in technologies that rely on flow-based detection, such as pyrosequencing or Roche 454 sequencing.⁸ In contrast, the Illumina GAIIX platform uses reversible, fluorescently labeled terminators, which allow each cycle to interrogate only one base at time, and thus, sequencing through homopolymer tracts on that platform is typically not a problem. In agreement, using the GAIIX, we were

able to accurately detect the correct call at the 2184delA position in over 99% of reads.

Although efforts were taken to ensure equimolar amplicon pooling, we still observed variations in read coverage. Similar coverage variability has been documented in other NGS platforms as well.⁹ Such variability may be attributed to differential adapter ligation during library preparation, preferential amplification during the emulsion PCR, enrichment efficiency, or sequencing bias. However, the G85E variant, which was under-represented consistently in the PGM data, also had relatively low coverage when sequenced on the Illumina platform. In this instance, the complexity of the G85E genomic region makes it a difficult target to sequence.

Following our experience with *CFTR* sequencing using the PGM, we believe several technical advances need to be addressed promptly before this technology is used widely in diagnostic-targeted, resequencing efforts. First, sample preparation remains relatively cumbersome and lengthy, making high-throughput sequencing difficult. Automation of the emulsion breaking and enrichment procedure using the One Touch instrument (Life Technologies) has reduced the hands on time. Second, to make PGM sequencing economically feasible, more barcodes need to be implemented. The ability to run more patients per chip will greatly reduce cost and turnaround time. Finally, improvements in homopolymer sequencing need to be made to reduce the number of false-positive calls. All but four of the 27 *CFTR* exons, including adjoining 30 bp of intronic junction regions, contain at least one homopolymer stretch of five or higher. Our data illustrate one consistent example where a false-deletion call in a *CFTR* 7-mer WT homopolymer stretch resulted in a false-positive mutant call. Full exon sequencing of the entire *CFTR* gene would likely reveal more calls, where actual mutant insertions are elevating the 5-mer or 6-mer homopolymer regions upward of seven. Also, in view of the fact that of the >1800 mutations in the *CFTR* database, >18% are frameshift mutations and insertion/deletions; although not all in homopolymer tracts, these mutation categories cannot be simply ignored.⁶ Whereas *CFTR* is one of the toughest gold standards for mutation detection, other genes do not have such high mutation rates. Regardless of the gene target, until improvements are made in homopolymer sequencing, a significant amount of confirmatory Sanger sequencing is required.

Despite these initial shortcomings in the PGM tech-

nology, we believe that there is high potential for the platform to be used in resequencing and genotyping efforts. With the use of a library representing all 23 *CFTR* mutations, we were able to identify all of the mutations following PGM sequencing. In a more realistic diagnostic test setting, PGM sequencing of a sample with two known CF disease-causing mutations correctly revealed both variants. Moreover, the speed and scalability of the PGM are optimal for diagnostic sequencing. The ability to generate data in several hours and not having to wait to batch samples for a run are attractive options. We have also been impressed with the robustness of the PGM, which requires very little maintenance, and as the technology relies on semiconductors rather than optics for detection, costly and time-consuming breakdowns are extremely rare.

In conclusion, we have shown initial feasibility of using the PGM platform to detect the 23 ACMG/ACOG *CFTR* mutations. With 316 and 318 chip improvements on the horizon and the rapid development of the Ion Torrent kits and pipeline, we are confident that the aforementioned shortcomings will be addressed in the immediate future. We look forward to the upcoming improvements that will establish the PGM as a robust platform in the clinical laboratory.

REFERENCES

1. Metzker ML. Sequencing technologies—the next generation. *Nat Rev Genet* 2010;11: 31–46.
2. Suzuki S, Ono N, Furusawa C, Ying B, Yomo T. Comparison of sequence reads obtained from three next-generation sequencing platforms. *PLoS One* 2011;6:e19534.
3. Voelkerding KV, Dames SA, Durtschi JD. Next-generation sequencing: from basic research to diagnostics. *Clin Chem* 2009;55: 641–658.
4. Pourmand N, Karhanek M, Persson HH, et al. Direct electrical detection of DNA synthesis. *Proc Natl Acad Sci USA* 2006;103: 6466–6470.
5. Pennisi E. Genomics. Semiconductors inspire new sequencing technologies. *Science* 2010;327:1190.
6. *Cystic Fibrosis Mutation Database*, <http://www.genet.sickkids.on.ca/cftr/>.
7. Watson MS, Cutting GR, Desnick RJ, et al. Cystic fibrosis population carrier screening: 2004 revision of American College of Medical Genetics mutation panel. *Genet Med* 2004;6:387–391.
8. Gilles A, Meglécz E, Pech N, Ferreira S, Malausa T, Martin JF. Accuracy and quality assessment of 454 GS-FLX titanium pyrosequencing. *BMC Genomics* 2011;12:245.
9. Dames S, Durtschi J, Geiersbach K, Stephens J, Voelkerding KV. Comparison of the Illumina Genome Analyzer and Roche 454 GS FLX for resequencing of hypertrophic cardiomyopathy-associated genes. *J Biomol Tech* 2010;21:73–80.